Utilisation de la couleur pour l'extraction de tableaux dans des images de documents[1]

Héloïse Alhéritière* – Florence Cloppet* – Camille Kurtz* – Nicole Vincent*

** Laboratoire LIPADE (EA 2517) – Université Paris Descartes*
45 rue des Saints Pères
75270 Paris Cedex 06
France
prénom.nom@parisdescartes.fr

RÉSUMÉ. *Les tableaux sont des éléments complexes qui peuvent perturber l'analyse automatique de la structure d'une image de document. Dans cet article, nous présentons une méthode fondée sur l'alternance de couleurs de lignes pour extraire des tableaux colorés à bordures non matérialisées. Les résultats expérimentaux obtenus à partir d'une base d'images de documents à mise en page variée, permettent de valider l'intérêt de cette approche.*

ABSTRACT. *Tables are complex elements that can disturb the automatic analysis of the structure of an image of a document. In this article, we present a method based on the alternation of the color of lines to extract color tables that are not materialized by physical rulings. Experimental results, obtained on a dataset of document images with various layouts, enable to validate the interest of this approach.*

MOTS-CLÉS : *Analyse d'images de documents, extraction de tableaux, détection de couleurs dominantes, segmentation d'images, croissance de régions.*

KEYWORDS: *Document image analysis, table extraction, dominant color detection, image segmentation, region growing.*

[1] Ce travail est financé par l'ANR (Agence nationale de la recherche française), dans le cadre du projet SHADES (ANR-14-CE28-0022).

1. Introduction

La génération et la numérisation de grandes masses de documents, mais également le souhait de les partager et de les diffuser rapidement, nécessitent de plus en plus la recherche et le développement d'outils informatiques permettant un accès efficace et pertinent à l'information contenue. Pour permettre un tel accès, une des étapes clés est de pouvoir analyser un document de manière automatique ou semi-automatique. Ce processus d'analyse d'images de documents est un processus complexe dû à la nature très hétérogène des contenus des documents. Pour analyser l'image d'un document, il est souvent nécessaire d'en rechercher au préalable sa structure : cette étape se fait généralement en cherchant ses caractéristiques ou ses règles de composition spatiale (régularité ou singularité des éléments le composant).

Depuis quelques années, l'extraction de la structure du document est une étape de plus en plus étudiée dans la littérature, en particulier dans la perspective de l'extraction d'informations, puis de connaissances, et qui pourront être traitées de manière automatisée. Différentes familles d'approches permettant de trouver la structure d'un document ont déjà été proposées : certaines cherchent à labéliser les zones d'un document après une première phase de segmentation d'images (Wang, Phillips, and Haralick 2006) tandis que d'autres cherchent à labelliser directement chaque pixel de l'image (Cote and Branzan Albu 2014). D'autres approches opèrent par une stratégie par couche cherchant à identifier chaque couche composant le document (comme la couche logo, imprimé ou tableaux). Nous nous intéressons dans cet article en particulier à l'extraction des tableaux à partir des images de documents qui, par leurs caractéristiques complexes, peuvent fortement perturber l'extraction des autres éléments composant la couche texte du document.

Les tableaux sont les éléments textuels les plus difficiles à définir. Selon le dictionnaire Larousse, un tableau est, dans le domaine mathématique, « un ensemble d'éléments disposés selon des lignes et des colonnes ou, de façon équivalente, dans les cases d'un rectangle quadrillé » et dans le domaine de l'imprimerie, qui est celui qui nous intéresse, un tableau est une « composition, encadrée ou non, comportant des chiffres et/ou des textes et divisée en colonnes ». Les tableaux ont ainsi des grandes variabilités dans leur mise en page (cf. Figure 1). Il n'y a aucune règle quant à la taille des cellules de ces derniers, ni sur les alignements qui peuvent être dans le même tableau à la fois au centre, à droite et à gauche. Les cellules fusionnées augmentent d'autant plus la difficulté de décrire ces éléments. Ce manque de règles rend particulièrement complexe l'extraction de tableaux sans bordures.

	colonne 1	colonne 2	colonne 3
ligne 1	12	texte	3,141
ligne 2	-36	nombre	2,718
ligne 3	0	29 euros	1,618

	colonne 1	colonne 2	colonne 3
ligne 1	12	texte	3,141
ligne 2	-36	nombre	2,718
ligne 3	0	29 euros	1,618

	colonne 1	colonne 2	colonne 3
ligne 1	12	texte	3,141
ligne 2	-36	nombre	2,718
ligne 3	0	29 euros	1,618

	colonne 1	colonne 2	colonne 3
ligne 1	12	texte	3,141
ligne 2	-36	nombre	2,718
ligne 3	0	29 euros	1,618

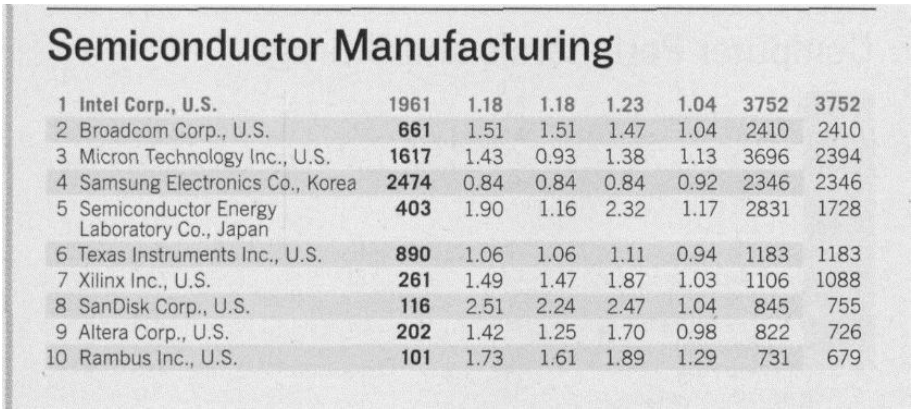
Figure 1: Exemples de tableaux à styles variés.

Plusieurs approches de détection et de localisation de tableaux ont été développées dans la littérature. Les stratégies implémentées par ces approches varient généralement en fonction d'*a priori* sur la façon dont les tableaux sont structurés au sein des images de documents. Les tableaux les plus simples à extraire, et également les tableaux les plus courants, sont ceux qui possèdent des séparateurs matérialisés par des traits et de nombreuses méthodes sont fondées sur cette caractéristique. Les auteurs de (Cesarini et al. 2002) utilisent ainsi la détection de lignes parallèles combinée à une approche par *MXY tree* pour localiser les tableaux. Une méthode proposée dans (Ramel et al. 2003) permet d'extraire le tableau en récupérant les lignes continues des documents. La méthode introduite dans (Gatos et al. 2005) repose sur un principe de détection de lignes horizontales et verticales des tableaux ainsi que de détection des points d'intersection.

Cependant, de plus en plus dans les documents, les tableaux ne sont plus forcément matérialisés par des traits (alternance de couleurs, espacements verticaux ou horizontaux, alignements, etc.). Les techniques d'extraction doivent alors reposer sur d'autres stratégies et caractéristiques pour détecter et localiser les tableaux. Par exemple, la méthode *T-Recs* (Kieninger and Dengel 2001) prend en entrée les mots segmentés puis s'intéresse aux alignements de ces derniers. (Mandal et al. 2006) utilisent l'espace régulier entre les colonnes. Les auteurs de (Shafait and Smith 2010) utilisent les *tab-stop* pour détecter les lignes pouvant faire partie d'un tableau et ainsi reconstituer ce dernier.

Dans ce contexte méthodologique et face à la complexité de ce problème d'analyse d'images, nous nous sommes intéressés dans ces travaux à la proposition d'une approche par texture pour extraire d'une image de document les tableaux possédant des lignes colorées, styles de tableaux qui sont de plus en plus utilisés dans la mise en page des documents (cf. Figure 2). Les tableaux à lignes colorées possèdent au moins deux couleurs de fond. Les couleurs des lignes sont ainsi alternées pour les différentier. Ils peuvent avoir d'autres couleurs qui souligneront l'importance d'une ligne comme une ligne d'en-tête ou de résultat. À l'opposé des

approches classiques reposant sur la détection de traits, l'approche que nous proposons se fonde sur la détection de l'alternance de lignes de couleurs dans une image de document.



Semiconductor Manufacturing

1	**Intel Corp., U.S.**	**1961**	**1.18**	**1.18**	**1.23**	**1.04**	**3752**	**3752**
2	Broadcom Corp., U.S.	**661**	1.51	1.51	1.47	1.04	2410	2410
3	Micron Technology Inc., U.S.	**1617**	1.43	0.93	1.38	1.13	3696	2394
4	Samsung Electronics Co., Korea	**2474**	0.84	0.84	0.84	0.92	2346	2346
5	Semiconductor Energy Laboratory Co., Japan	**403**	1.90	1.16	2.32	1.17	2831	1728
6	Texas Instruments Inc., U.S.	**890**	1.06	1.06	1.11	0.94	1183	1183
7	Xilinx Inc., U.S.	**261**	1.49	1.47	1.87	1.03	1106	1088
8	SanDisk Corp., U.S.	**116**	2.51	2.24	2.47	1.04	845	755
9	Altera Corp., U.S.	**202**	1.42	1.25	1.70	0.98	822	726
10	Rambus Inc., U.S.	**101**	1.73	1.61	1.89	1.29	731	679

Figure 2 : Exemple d'un tableau à bordures non matérialisées et possédant une alternance de couleurs entre les lignes.

Cet article s'articule de la manière suivante. Nous présentons en section 2 la méthode proposée pour l'extraction de tableaux avec une alternance de couleurs. En section 3, nous proposons une étude expérimentale préliminaire permettant d'évaluer la pertinence de cette méthode sur une base d'images de documents contenant de nombreux types de tableaux. Nous terminons en section 4 par des discussions et un ensemble de perspectives.

2. Méthode proposée pour l'extraction de tableaux

La méthode proposée permet d'extraire dans une image de document des tableaux possédant des lignes colorées avec alternance. Elle prend en entrée une image de document en couleur dans l'espace RGB et fournit en sortie une information sur le nombre de tableaux dans le document ainsi que la localisation spatiale de ces derniers au sein de l'image. Les différentes étapes de l'approche proposée sont résumées en Figure 3.

Les tableaux étant des éléments textuels et ainsi, le plus souvent, composés en majorité de mots, la détection des tableaux débute par une première étape (cf. Section 2.1) d'extraction des mots à partir de l'image du document. Cette approche étant dédiée aux tableaux avec alternance de couleurs, la deuxième étape (cf. section 2.2) consiste à déterminer les couleurs dominantes de fond des mots du document puis à classifier l'ensemble des mots en fonction de cette caractéristique. Une fois les

mots de couleurs dominantes identifiés, la dernière étape (cf. section 2.3) consiste à recomposer le tableau en se fondant sur : la reconstruction des cellules, *via* une opération de segmentation d'images, puis la reconstruction des lignes colorées, et enfin sur l'alternance de ces lignes colorées de manière à obtenir la structure tabulaire.

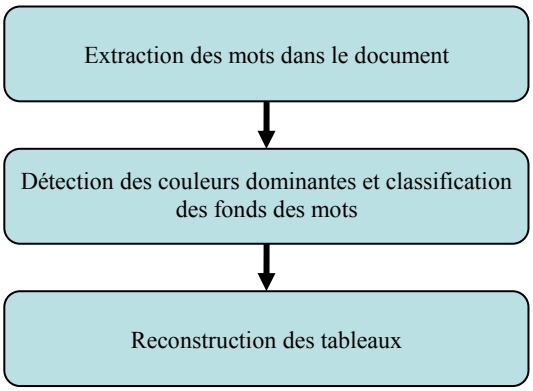


Figure 3: Étapes de la méthode d'extraction des tableaux.

2.1 *Extraction des mots du document*

Étant donnée une image de document, la méthode utilisée pour extraire l'ensemble des mots qui composent le document repose sur une étape préalable de binarisation d'image. Une fois cette binarisation réalisée, les composantes connexes sont extraites par une étape de filtrage sur un critère de taille permettant de récupérer les caractères constituant le document. À partir des caractères extraits, une dernière étape permet de reconstruire les mots du document *via* un critère d'alignement horizontal.

Après cette première étape d'extraction des mots de l'image de document, on dispose de l'ensemble du contenu textuel du document (d'un point de vue pixels) ainsi que de la location spatiale des mots du document (en particulier des coordonnées (x, y) des quatre coins de la boite englobante de chaque mot).

2.2 *Détection des couleurs dominantes et classification des fonds des mots*

Notre approche se focalisant sur l'extraction de tableaux avec une alternance de couleurs, il est au préalable nécessaire d'identifier ces différentes

couleurs, que nous qualifierons de couleurs « dominantes » au sein de la partie textuelle du document. Les couleurs que nous cherchons à identifier sont les couleurs du fond des mots extraits à l'étape précédente.

Pour ce faire l'approche proposée commence par extraire le fond de chaque mot (pour chaque mot, séparation des pixels de texte vs. des pixels de fond). L'identification des couleurs dominantes utilisées dans le document se fait alors par la classification des fonds des mots.

(a) 

(b) 

(c) 

Figure 4 : Imagette du mot « Telecom » : (a) initiale ; (b) après binarisation ; (c) fond extrait.

La méthode de classification proposée consiste à rechercher récursivement les n couleurs dominantes des fonds des mots du document. Pour ce faire, nous utilisons un histogramme modélisant la distribution des couleurs RGB de toutes les images des fonds des mots extraits du document. Une image de fond de mot correspond ici à une imagette de la boite englobante du mot à laquelle on retranche la partie textuelle (obtenue à l'étape précédente) pour ne conserver que les pixels composant le fond du mot (cf. Figure 4).

À partir du mode M de l'histogramme, la stratégie est de déterminer un intervalle I centré en M d'amplitude μ. Une fois cet intervalle déterminé, il s'agit ensuite d'identifier tous les mots dont les fonds possèdent un pourcentage de pixels ayant une couleur dans l'intervalle I supérieure ou égale à un seuil S fixé de manière empirique. Le paramètre μ autorise ainsi une certaine tolérance vis à vis de variations de couleurs dues aux bruits résultant par exemple de l'étape de segmentation. Les mots ainsi récupérés sont regroupés par couleur de fond formant ainsi une classe de mots de couleur dominante. L'histogramme est ainsi recalculé en y retranchant les mots identifiés lors de cette étape. Cette étape est appliquée récursivement jusqu'à ce qu'il ne soit plus possible d'identifier un mode M significatif (relativement à la recherche d'un intervalle I centré en M d'amplitude μ).

La méthode fournie en sortie les *n* couleurs majoritaires des fonds des mots du document ainsi qu'une classification en *n* classes des mots du document. Une fois ces résultats obtenus, l'étape suivante consiste reconstruire le tableau en se fondant sur l'alternance de lignes entre couleurs.

2.3 *Reconstruction des tableaux*

Pour reconstruire les tableaux à alternance de couleurs, la stratégie consiste ici à considérer une des caractéristiques principales (généralement invariante) structurant les tableaux : chaque ligne possède des cheminées permettant de distinguer les cellules, c'est-à-dire un espace continu verticale entre les lignes.

La classification effectuée à l'étape précédente a permis d'obtenir le nombre *n* de couleurs majoritaires des fonds des mots du document. Cependant, il est à noter que ces *n* couleurs ne représentent pas forcément que les couleurs des lignes alternées des tableaux (en plus de la couleur principale de fond des mots du document, en général le blanc). En effet, le reste du document peut être composé de mots surlignés avec des couleurs afin de les mettre en valeur. Par conséquent, l'extraction des couleurs majoritaires ne fournit pas directement l'emplacement des tableaux. Par ailleurs, des défauts relatifs à la luminosité lors de l'acquisition de l'image du document ou à un fond non uniforme ont pu induire des erreurs lors de l'étape de détection des couleurs et de classification des mots : ces erreurs pourront être corrigées lors de la reconstruction des tableaux.

L'étape de reconstruction des tableaux contient plusieurs sous-étapes (cf. Figure 5) : à partir des mots séparés en *n* classes de couleur, la première sous-étape consiste à reconstruire les cellules ; à partir de ces dernières, une deuxième étape permet la reconstruction des lignes du tableau. La dernière sous-étape consiste alors à reconstruire le tableau. Pour ce faire, l'idée est de rechercher dans les lignes retrouvées celles qui peuvent former un tableau *via* une analyse de leur alternance relativement aux *n* couleurs dominantes.

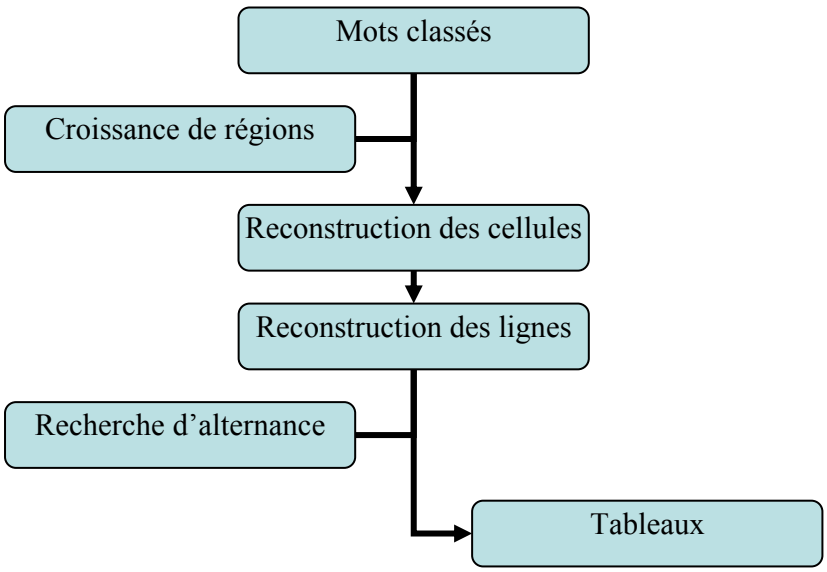


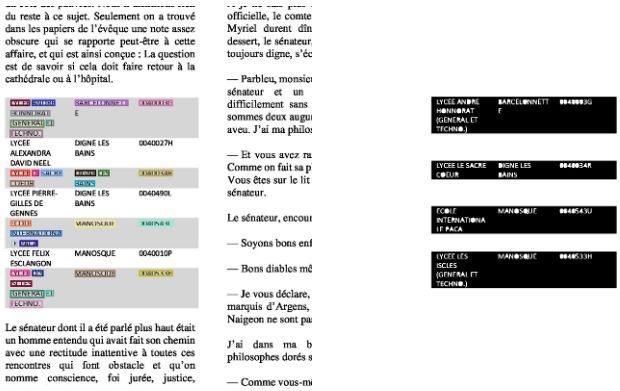
Figure 5: Sous-étapes de l'approche de reconstruction des tableaux.

Les *n* classes de mots de couleur sont triées de manière croissante en fonction du cardinal de chaque classe. La méthode est initialisée par la classe *i* de mots de couleur possédant le plus petit cardinal (puis répétée avec la classe suivante, etc.). Nous faisons ici l'hypothèse que cette couleur est l'une des couleurs utilisées pour l'alternance des lignes dans le tableau. Généralement, les mots constituant un tableau sont minoritaires dans l'ensemble du document.

Reconstruction des cellules Pour chaque mot de la classe *i*, le principe est de reconstituer la cellule contenant ce mot en utilisant une méthode de segmentation par croissance de région (*region growing)*. Les graines de cette dernière sont initialisées avec des points connexes extérieurs à ceux composant le rectangle englobant du mot et en utilisant un critère de croissance lié à l'homogénéité par rapport à la moyenne de la classe (cf. Figure 6). Le critère d'arrêt de la croissance est fonction du paramètre μ *(*utilisé dans l'étape de détection et de classification des couleurs, cf. section 2.2) déterminant l'amplitude de l'intervalle pour sélectionner le mode de l'histogramme.

Il se peut que le tableau, à laquelle la cellule courante appartient, possède des bordures de la même couleur que le fond du mot courant. Dans ce cas précis, la méthode de croissance de région peut former des composantes connexes s'étendant dans les traits du tableau. La cellule ainsi reconstruite pourra alors potentiellement contenir tout le tableau. Ce cas est alors détecté en analysant, dans la boite

englobante de la cellule ainsi formée, le pourcentage de mots n'appartenant pas à la couleur de la classe *i*. Dans certains cas spécifiques, quand les cellules ne sont pas isolées par des séparateurs matérialisés, la reconstruction d'une cellule permettra directement de reconstruire la ligne entière du tableau à laquelle la cellule appartient (cf. Figure 6).



(a) (b)

Figure 6: (a) les mots de la classe minoritaire sont entourés ; (b) les résultats de la croissance de région à partir des mots entourés dans l'image de gauche.

Reconstruction des lignes Une fois les cellules contenant les mots de la classe *i* reconstituées, la prochaine étape consiste à reconstruire les lignes contenant ces cellules. Toutes les cellules d'un tableau ne contiennent pas forcément des mots (cf. Figure 7 (a)) et certains mots à fonds colorés ont pu être mal classés par l'étape de classification. Pour faire face à ce problème, et ainsi reconstituer la ligne entière, l'approche proposée consiste à vérifier sur les côtés (à gauche et à droite) d'une cellule reconstruite s'il n'y a pas d'autres cellules non reconnues.

Les cellules alignées horizontalement peuvent faire partie d'une même ligne ou de deux tableaux différents. Pour vérifier cela, il convient de s'assurer qu'il y a bien une continuité entre toutes les cellules. La stratégie est ici de réutiliser un algorithme de croissance de région pour étendre les cellules en initialisant les graines de l'algorithme avec des points se trouvant à gauche ou à droite des limites de la cellule (cf. Figure 7 (b, c)). En réappliquant cette stratégie à toutes les cellules contenant les mots de la classe *i*, on obtient les lignes du tableau contenant les cellules de cette classe (cf. Figure 7 (d)).

(a) **(b)** **(c)** **(d)**

Figure 7 : Sous-étapes considérées lors de la reconstruction des cellules et des lignes du tableau : (a) les mots de la classe de plus petit cardinal ; (b) croissance de régions sur les mots pour reconstruire les cellules ; (c) croissance de régions sur les cellules pour reconstruire les lignes ; (d) les lignes reconstruites.

Reconstruction des tableaux La dernière sous-étape de la reconstruction consiste à vérifier si les lignes reconstruites précédemment peuvent former un tableau via une analyse de leur alternance, relativement aux couleurs dominantes.

L'idée est d'analyser l'alternance des lignes reconstituées formées des mots de la classe *i* par rapport aux lignes reconstituées avec les mots de la classe *j*. Pour que 2 lignes consécutives de la classe *i* fassent partie d'un tableau, il faut vérifier :

- que celles-ci ont une intersection non-vide sur l'axe des abscisses (par projection sur l'axe X) ;
- qu'il existe des mots faisant partie de la classes *j* entre ces deux lignes, et que ces mots laissent apparaitre une cheminée ;
- qu'il n'existe pas des mots d'autres classes (que *i* et *j*) entre ces lignes.

Ce processus de reconstruction de tableau est appliqué sur toutes les couleurs excepté la couleur majoritaire (liée à la classe de plus grand cardinal) car il peut y avoir plusieurs tableaux avec des couleurs différentes dans le même document.

3. Etude expérimentale

3.1. *Jeu de données*

La méthode proposée a été évaluée sur une base de données comportant soixante-quinze images de documents, de résolution 200dpi, avec du texte sur deux colonnes. Au sein de cette base, 15 images de documents ne comportent pas de tableaux. Les 60 images restantes comportent des tableaux avec des lignes de couleurs alternées. Ces images peuvent avoir jusqu'à 4 couleurs dans le document et de 2 à 3 couleurs dans les cellules des tableaux. Quelques images issues de cette base d'images sont présentées dans la Figure 8.

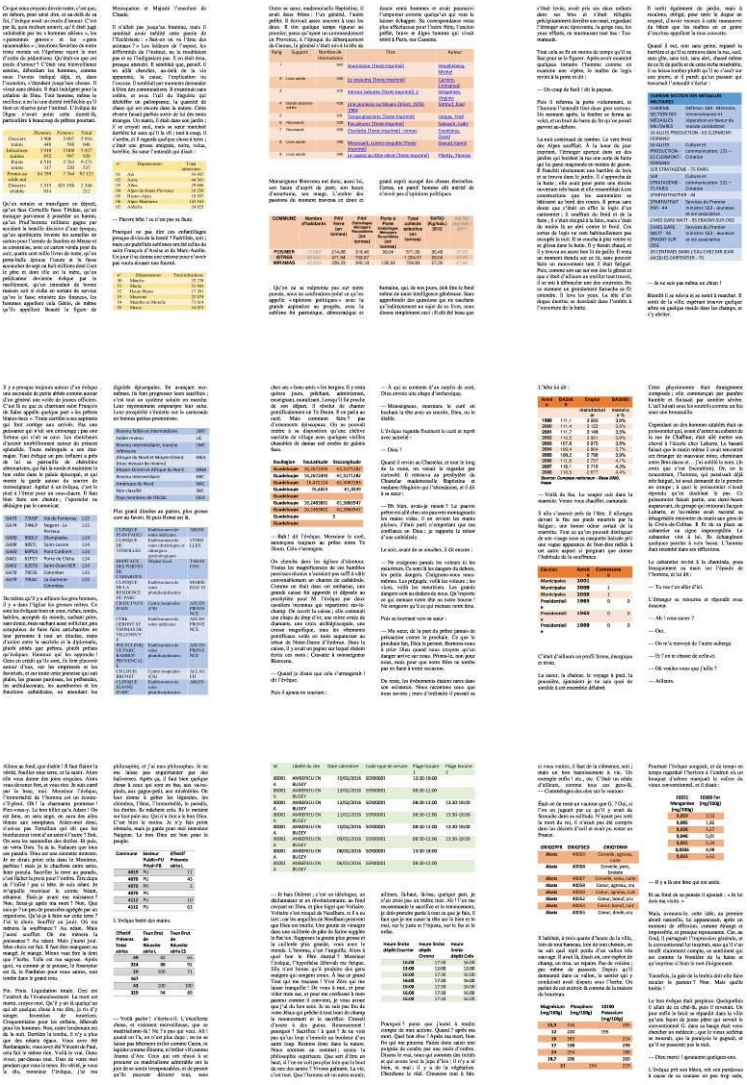


Figure 8 : Exemples d'images de documents composant le jeu de données.

3.2. *Protocole expérimental*

La méthode a été implémentée en C++ via la bibliothèque d'analyse d'images Leptonica[2]. Les paramètres de la méthode ont été fixés de manière expérimentale. Dans cette première étude, nous avons fixé les paramètres de façon à maximiser les scores d'extraction des tableaux. Les paramètres liés à la détection de couleurs (cf. section 2.2) sont fixés empiriquement avec μ égal à 15 et S égal à 60%. Le même μ est utilisé pour la partie de reconstruction des cellules (cf. section 2.3) pour le critère d'arrêt de l'algorithme de croissance de région.

Pour chaque image de la base, nous disposons d'une vérité terrain, c'est-à-dire qu'une boite englobante est dessinée autour des tableaux contenus dans chaque image. Nous disposons également pour chaque image du nombre de tableaux présents (ou non) dans l'image de document. Ces informations nous permettent d'évaluer qualitativement (visuellement) et quantitativement les résultats fournis par l'approche proposée pour l'extraction de tableaux.

Dans le cadre de ce protocole expérimental, nous évaluons indépendamment la capacité de notre méthode pour la détection de tableaux dans un document (tableau présent ou non, ainsi que le nombre de tableaux présents) et pour la localisation spatiale des tableaux au sein des images (intersection entre la boite englobante du tableau localisé et la boite englobante du tableau dans la vérité terrain). Pour évaluer quantitativement la détection et la localisation des tableaux, nous avons calculé les indices de précision et de rappel.

3.3. *Présentation et analyse des résultats*

La Figure 9 présente quelques résultats d'extraction de tableau à lignes colorées alternées via la méthode que nous proposons dans ces travaux. Ces résultats permettent d'apprécier visuellement l'intérêt et certaines limites de la méthode proposée. En particulier, on peut remarquer que même en cas de tableaux multiples au sein d'un même document, ces derniers sont correctement localisés. Cependant, on remarque également que la méthode proposée ne parvient pas à reconstruire au sein des tableaux la ligne d'en-tête de ces derniers lorsque celle-ci n'est pas de couleur minoritaire.

La table 1 présente les scores obtenus relatifs à la détection des tableaux au sein des documents. Nous avons choisi ici de regrouper les documents du jeu de données en 3 catégories : documents sans tableau, documents ne comportant qu'un tableau, documents comportant plusieurs tableaux.

On observe ici que la méthode, pour les 3 catégories de documents, détecte correctement si un ou plusieurs tableaux sont présents (ou non). Lorsque plusieurs

[2] http://www.leptonica.com/

tableaux sont présents, certains tableaux ne sont pas détectés. Ceci est dû aux cas où deux tableaux sont présents sur une même colonne et qu'ils ne sont séparés que par une ligne unique : dans une telle configuration la méthode proposée ne détecte qu'un tableau à la place de deux.

La localisation a montré que si les tableaux étaient globalement correctement détectés, la première et la dernière ligne ne l'étaient pas toujours, sauf dans le cas où les deux couleurs utilisées pour les tableaux sont différentes de la couleur de fond.

Pour évaluer la qualité de l'extraction des tableaux, nous utilisons les critères suivants :

— Le nombre de vrais positifs VP : pixels trouvés comme appartenant au tableau et appartenant au véritable tableau ;

— Le nombre de faux négatif FN : nombres de pixels trouvés comme n'appartenant pas au tableau et appartenant au véritable tableau ;

— Le nombre de faux positifs FP : pixels trouvés comme appartenant au tableau et n'appartenant pas au véritable tableau ;

— Le nombre de vrai négatifs VN : nombres de pixels trouvés comme n'appartenant pas au tableau et n'appartenant pas au véritable tableau ;

Nous évaluerons la précision et le rappel grâce aux formules suivantes :

$$Precision = \frac{VP}{VP + FP} \qquad Rappel = \frac{VP}{VP + FN}$$

Les tables 2 et 3 présentent les valeurs des scores relatifs à la localisation des tableaux au sein des documents. Nous avons choisi ici de regrouper les documents du jeu de données de deux façons différentes : relativement au nombre de tableaux dans les documents (documents sans tableau, documents ne comportant qu'un tableau, documents comportant plusieurs tableaux) et relativement au nombre de couleurs utilisées pour composer les documents (sans tableau, avec 2, 3 ou 4 couleurs dont 2 ou 3 pour les tableaux).

Si les tableaux sont correctement détectés, il manque souvent la première ou la dernière ligne, ce qui implique un rappel plus faible que la précision.

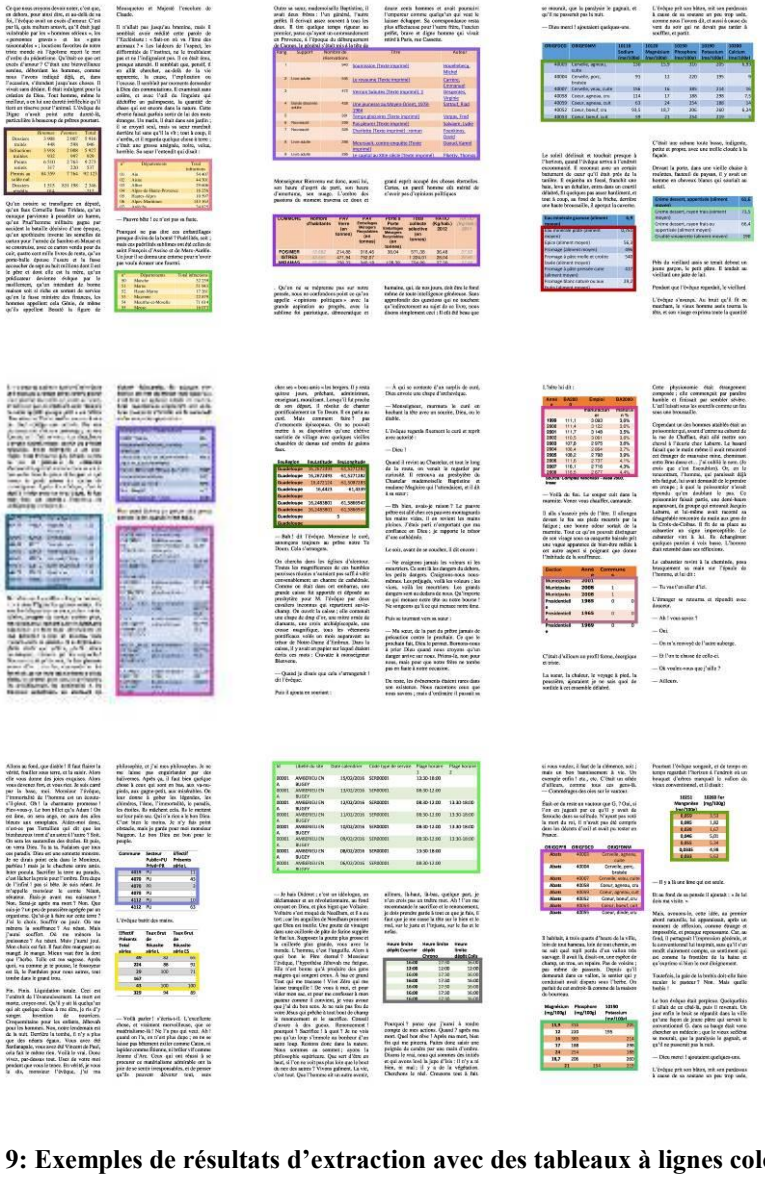


Figure 9: Exemples de résultats d'extraction avec des tableaux à lignes colorées alternées.

Catégories de documents	Précision	Rappel
Sans tableau (15 doc.)	100%	100%
Avec un tableau (20 doc.)	100%	100%
Avec plusieurs tableaux (40 doc.)	98%	100%

Table 1 : Scores (rappel et précision) obtenus pour la détection des tableaux dans la base de données expérimentale.

Catégories de documents	Précision	Rappel
Sans tableau (15 doc.)	100%	100%
Avec un tableau (20 doc.)	100%	89%
Avec plusieurs tableaux (40 doc.)	98%	83%

Table 2: Scores (rappel et précision) obtenus pour l'extraction des tableaux dans la base de données expérimentale (regroupement par nombres de tableaux dans les documents).

Catégories de documents	Précision	Rappel
Sans tableau (15 doc.)	100%	100%
Avec 2 couleurs dont 2 pour les tableaux (15 doc.)	97%	81%
Avec 3 couleurs dont 2 pour les tableaux (15 doc.)	99%	99%
Avec 3 couleurs dont 3 pour les tableaux (15 doc.)	100%	80%
Avec 4 couleurs dont 3 pour les tableaux (15 doc.)	98%	81%

Table 3 : Scores (rappel et précision) obtenus pour l'extraction des tableaux dans la base de données expérimentale (regroupement par nombre de couleurs dans les documents).

4. Discussions et perspectives

Dans cet article, nous avons introduit une nouvelle approche utilisant la couleur pour l'extraction de tableaux dans des images de documents. Par rapport aux approches classiques fondées sur l'extraction de traits, la principale originalité de cette méthode repose sur le fait de considérer l'alternance de couleurs de lignes pour extraire des tableaux à bordures non matérialisées. Les premiers résultats obtenus sur une base de documents à mise en page variée montrent que la reconnaissance de tableaux complexes au sein d'images de documents, peut être fondée sur cette caractéristique.

Ces travaux ouvrent la voie à différentes perspectives. Si la méthode présentée permet de retrouver un tableau dématérialisé, elle ne permet pas de

retrouver la dernière ligne si celle-ci ne fait pas partie de la couleur minoritaire analysée.

Un post-traitement pourra également être utilisé pour récupérer la troisième couleur si celle-ci fait partie de l'en-tête ou de la colonne d'en-tête. Par ailleurs, cette méthode pourra être étendue afin de trouver les tableaux avec des colonnes aux couleurs alternées.

5. Bibliographie

Cesarini, Francesca, Simone Marinai, L Sarti, and Giovanni Soda. 2002. "Trainable Table Location in Document Images." In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 236–40.

Cote, Melissa, and Alexandra Branzan Albu. 2014. "Texture Sparseness for Pixel Classification of Business Document Images." *International Journal on Document Analysis and Recognition (IJDAR)* 17 (3): 257–73.

Gatos, Basilis G., Dimitrios Danatsas, Ioannis Pratikakis, and Stavros J. Perantonis. 2005. "Automatic Table Detection in Document Images." In *Proceedings of the International Conference on Pattern Recognition and Data Mining (ICPRDM)*, 3686:609–18.

Kieninger, Thomas, and Andreas Dengel. 2001. "Applying the T-Recs Table Recognition System to the Business Letter Domain." In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*.

Mandal, S, S P Chowdhury, a K Das, and Bhabatosh Chanda. 2006. "A Simple and Effective Table Detection System from Document Images." *International Journal on Document Analysis and Recognition (IJDAR)* 8 (2-3): 172–82.

Ramel, Jean-Yves, Michel Crucianu, Nicole Vincent, and Claudie Faure. 2003. "Detection, Extraction and Representation of Tables." In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 374–78.

Shafait, Faisal, and Ray Smith. 2010. "Table Detection in Heterogeneous Documents." In *Proceedings of the IAPR International Workshop on Document Analysis Systems (DAS)*, 65–72.

Wang, Yalin, Ihsin T. Phillips, and Robert M. Haralick. 2006. "Document Zone Content Classification and Its Performance Evaluation." *Pattern Recognition (PR)* 39: 57–73.